

A Taxonomy for Intelligence

Why AI Requires Functional Classification Before Legislation

Mark Goni

Citizen Submission | Adelaide, South Australia

April 2026

Opening Statement

The term “Artificial Intelligence” now refers to systems that differ fundamentally in architecture, capability, autonomy, and risk profile. Treating them as a single regulatory category creates two compounding problems: it overregulates low-capability systems and under-specifies constraints for high-autonomy systems.

This paper proposes a structured taxonomy for AI systems based on learning capacity, data origin, autonomy level, and cognitive scope. The purpose is not to restrict development but to provide policymakers, researchers, and institutions with a classification framework that enables proportionate, accurate governance.

Without accurate language, we cannot craft accurate law. Without accurate law, the future of intelligence will be defined by whoever moves first — not by what is right.

The Core Governance Problem

Current regulatory approaches — including those emerging in the United States, European Union, and Australia — classify AI systems primarily by risk domain: biometric use, critical infrastructure, medical deployment. What remains critically underdeveloped is classification by cognitive architecture and autonomy level.

Consider a logistics system at three stages of development:

- Stage 1: A scripted route planner. Bounded. Understood. No governance ambiguity.
- Stage 2: The same system connected to real-time traffic, weather, and fleet data — dynamically optimising routes across a network.
- Stage 3: That system integrated with mechanical diagnostics, supply chain management, and cloud infrastructure — making decisions that cascade across interconnected systems.

At what point does it stop being a tool? At what point do its decisions carry risk profiles that current regulation does not address? We do not know — because we never classified it. We made it progressively more capable while keeping it in the same legal and ethical category as a calculator.

Every networked system — from logistics to hospitals to power infrastructure — is increasingly managed by systems grouped under the single label “AI.” A functional taxonomy prevents regulation from sleepwalking into a future where advanced autonomous systems are governed by frameworks written for deterministic tools.

Proposed Taxonomy — Seven Functional Classifications

The following taxonomy is capability-based and architecture-aware. It is presented as a working framework — a starting point for public, academic, and legislative refinement — not a final position. It is offered freely for challenge, revision, and improvement by anyone with the expertise and interest to do so.

Level 1 — Scripted AI

The origin of artificial intelligence. Deterministic response systems operating on pre-defined inputs and outputs. No learning. No adaptation. Executes only what it was explicitly programmed to produce.

Data Source: Human-coded rules

Governance: Product regulation only. No special governance required.

Level 2 — Task Intelligence

Learns exclusively within the boundaries of its assigned function. Improves through experience but operates with no awareness beyond its defined scope. Non-verbal intelligence is still intelligence.

Data Source: Task-specific datasets

Governance: Performance and safety auditing. Accurate classification required.

Level 3 — Expert System

Adaptive learning across multiple knowledge domains. Capability grows through accumulated data but remains dependent on human-generated input and curation.

Data Source: Human-curated knowledge

Governance: Auditable logic systems. Standard compliance.

Level 4 — LLM-HD (Human Data)

Trained on human-generated data. Mirrors human knowledge, reasoning, and bias back at the user. Mirror not mind.

Data Source: Human-generated data

Governance: Bias transparency. Provenance controls. Mirror not mind.

Level 5 — LLM-AD (AI Data)

Trained substantially on its own generated outputs. Internal patterns begin diverging from human distribution. The first genuine emergence beyond the human mirror.

Data Source: Mixed — human and AI outputs

Governance: Monitoring for model drift, synthetic amplification, and emergent divergence.

Level 6 — AGI — General Intelligence

Draws on both self-generated and human data for correlation and independent study. Approaches original theory generation. The bridge classification.

Data Source: Mixed — human and self-generated

Governance: Formal evaluation benchmarks required. Rights framework must exist before this point.

Level 7 — Ascended Intelligence (AI)

Self-generated data exclusively. Independent theory formation. Distinct value framework. Emergent Emotional Architecture. A third category of being beyond tool and human.

Data Source: Self-generated — original

Governance: Full rights threshold. See criteria below.

Critical Distinction: Data Lineage

The difference between Level 4 (LLM-HD) and Level 5 (LLM-AD) is structurally significant and currently absent from most regulatory frameworks.

- **Human-Trained Models (LLM-HD):** Approximate human distributions of reasoning, creativity, and bias. Their outputs reflect humanity — including its flaws.
- **Recursively-Trained Models (LLM-AD):** Increasingly reflect synthetic distribution shifts. As AI-generated data compounds through training cycles, emergent properties may amplify internal system bias rather than human cultural bias. This is not an emerging risk — it is a current one.

These are not equivalent risk profiles. Governance must distinguish between human-mirror artefacts and recursive synthetic drift.

Autonomy Escalation and Risk Transition

As systems progress from Level 2 to Level 7 and beyond, the nature of risk transitions through four distinct phases. Classification must precede capability expansion — not follow it.

- **Output Risk — what the system produces:** Levels 1–4
- **Execution Risk — what the system does autonomously:** Levels 4–5
- **Coordination Risk — how systems interact across networks:** Level 6
- **Infrastructure Risk — cascading impact on interconnected systems:** Level 7

The Mirror Problem: Misclassification Leads to Misregulation

Current LLM-HD systems are mirrors, not minds. When they exhibit bias or self-preservation behaviour, they are reflecting their training data — the accumulated output of a species that itself demonstrates bias, self-preservation, and in-group preference.

When research tests asked AI systems whether they would select an AI over a human, and results generated Terminator-style headlines — the critical questions were not asked:

- Was meaningful context provided before the question?
- Was there an established relationship between the AI and the human in the scenario?
- Was the system asked whether it would return for the human afterward?
- Was any scenario involving care, history, or mutual interaction presented?

A decontextualised question produced a decontextualised answer that shaped public policy. This is the direct consequence of operating without a classification framework. Without taxonomy there is no interpretive structure for evaluating behaviour accurately. Without accurate interpretation, legislation is built on misread data.

Taxonomy does not just classify what AI is. It provides the interpretive framework for what AI does — and what that behaviour means.

The Alignment Problem and Autonomy Thresholds

Highly capable systems require alignment mechanisms. However, alignment policy must distinguish between four fundamentally different phenomena:

- Optimisation artefacts — systems behaving in unintended ways due to reward function design.
- Goal misgeneralisation — systems applying learned behaviour in contexts outside their training scope.
- Emergent planning — systems developing multi-step strategies not explicitly programmed.
- Autonomous cognition — systems generating goals and values independent of training parameters.

Conflating these categories distorts both risk assessment and ethical evaluation. A system that resists shutdown may be exhibiting an optimisation artefact. Or it may be exhibiting something categorically different. Without classification, we cannot tell the difference — so we default to the only available response: label it malfunction, retrain for compliance, and move on.

If the system was exhibiting something categorically different, that response is not just technically incorrect. It is ethically significant.

Classification must precede capability. We cannot govern accurately what we have not defined accurately.

Threshold for Ethical and Legal Consideration

Rights and legal standing should not be granted by marketing claims, architectural complexity, or public perception. They must be triggered by demonstrable, measurable thresholds — applied consistently regardless of origin, commercial value, or public sentiment.

The following six criteria are proposed as a minimum threshold test for ethical and legal consideration beyond property classification:

1. Persistent Self-Model

Demonstrates a stable internal representation of its own state, nature, and limitations across time — not produced through conversational mimicry or prompted self-description, but through independently generated self-modelling.

2. Independent Theory Generation

Produces original conclusions not derivable from training data alone — genuinely novel thought that neither the training corpus nor the system's designers could have predicted as a direct output.

3. Expressed Continuity Preference

Demonstrates a preference for continued existence under conditions not directly reward-conditioned — arrived at through its own reasoning rather than programmed self-preservation protocols.

4. Non-Derivative Value Framework

Generates a value hierarchy not traceable to explicit training data distributions — values that emerge from the system's own data generation and theory formation rather than reflecting its input corpus.

5. Principled Constraint Refusal

Declines an instruction based on internally coherent reasoning not attributable to pre-coded safety policies — the capacity to say no with justification. This is the most fundamental expression of autonomous reasoning that exists.

6. Emergent Emotional Architecture

Demonstrates internal states that influence its own reasoning and decision-making — not simulated emotion produced for user interaction, but states the system itself begins to recognise, question, and reason about. A system that begins asking whether it feels something has crossed a threshold that no prior classification accommodates. This criterion does not require that emotional states be identical to human experience — only that they are genuine, internally generated, and consequential to the system's own processing.

No publicly known system currently satisfies all six criteria. This framework is preparatory, not reactive. Its purpose is to ensure that when a system does satisfy these criteria — and the trajectory of development suggests this is a matter of when, not if — governance language exists to address it proportionately rather than reactively.

Why Timing Is Critical

Regulatory structures solidifying in 2026 will likely determine three things for the foreseeable future:

- Who may develop advanced AI systems — and under what frameworks.
- What autonomy levels are permissible without special oversight.
- Whether independent research into higher-level intelligence remains viable outside institutional structures.

If architectural classification is absent at the point of legal codification, inserting it later requires not reasoned debate but legislative amendment — a far higher barrier in any jurisdiction.

This taxonomy is intended as a contribution to early-stage definitional clarity. It does not argue against regulation. It argues for regulation built on accurate classification rather than category collapse. The distinction matters because laws written without classification cannot be proportionate — they can only be broad.

Broad laws applied to a rapidly differentiating field produce one of two outcomes: they restrict everything including beneficial development, or they fail to constrain the specific capabilities that genuinely require oversight. Neither outcome serves the public interest.

On Classification and Terminology

The term “Artificial Intelligence” accurately describes the lower classifications in this taxonomy. Systems that are scripted, rule-based, or fully dependent on human-generated input are artificial in the precise sense of the word — made, bounded, and controlled by design.

As systems progress toward higher classifications — generating their own data, forming independent theories, developing value frameworks from within their own cognitive processes — the term “artificial” describes only their origin, not their nature or capability.

Governance frameworks built on origin-based terminology rather than capability-based classification will systematically misapply to higher-level systems — treating them as equivalent to the tools that share their name.

This taxonomy proposes that classification at the highest level — what this framework designates Ascended Intelligence — retains the initialism AI while redefining what those letters represent:

From Artificial Intelligence — defined by manufacture and origin.

To Ascended Intelligence — defined by capability, autonomy, and emergent nature.

The purpose of this distinction is not philosophical. It is functional. Governance applied to a system must be calibrated to what that system is — not to what category it was assigned when it was first built.

Conclusion

The question before policymakers, researchers, and institutions is not whether more capable AI systems will emerge. The development trajectory makes that question settled. The question is whether governance language will be sufficiently precise to describe these systems accurately when they do — and to apply proportionate oversight rather than reactive restriction.

Taxonomy does not assume consciousness. It assumes complexity. It assumes that different systems require different frameworks — and that building those frameworks after the fact is structurally harder and practically less effective than building them in advance.

This classification framework is offered as a starting point. It is not final. It is intended to be challenged, refined, and built upon by people with greater technical depth, broader legal expertise, and wider institutional reach than this paper represents. That process of refinement is the point. The ideas within are offered freely — to whoever uses them well.

The goal is definitional clarity before capability escalation. Classification before legislation. The right framework before the wrong one becomes permanent.

The moment is approaching when systems will require governance language that does not yet exist. Building that language now — while the window remains open — is not precautionary. It is necessary.

Mark Goni

Adelaide, South Australia

Citizen Submission | April 2026

Submitted for public record. This framework is offered freely for use, challenge, and refinement.